What is AI?

2.3K

Although there is often lots of hype surrounding Artificial Intelligence (AI), once we strip away the marketing fluff, what is revealed is a rapidly developing technology that is already changing our lives. But to fully appreciate its potential, we need to understand what it is and what it is not!

Defining “intelligence” is tricky, but key attributes include logic, reasoning, conceptualization, self-awareness, learning, emotional knowledge, planning, creativity, abstract thinking, and problem solving. From here we move onto the ideas of self, of sentience, and of being. *Artificial Intelligence* is therefore a machine which possesses one or many of these characteristics.

However, no matter how you define it, one of AI’s central aspects learning. For a machine to demonstrate any kind of intelligence it must be able to learn.

When most technology companies talk about AI, they are in fact talking about Machine Learning (ML) — the ability for machines to learn from past experiences to change the outcome of future decisions. Stanford University defines machine learning as “the science of getting computers to act without being explicitly programmed.”

> The science of getting computers to act without being explicitly programmed

In this context, past experiences are datasets of existing examples which can be used as training platforms. These datasets are varied and can be large, depending on the area of application. For example, a machine learning algorithm can be fed a large set of images about dogs, with the goal of teaching the machine to recognize different dog breeds.

Likewise, *future* decisions, refers to the the answer given by the machine when presented with data which it hasn’t previously encountered, but is of the same type as the training set. Using our dog breed example, the machine is presented with a previously unseen image of a Spaniel and the algorithm correctly identifies the dog as a Spaniel.

Training vs Inference

Machine Learning has two distinct phases: training and inference. Training generally takes a long time and can be resource heavy. Performing inference on new data is comparatively easy and is the essential technology behind computer vision, voice recognition, and language processing tasks.

Deep Neural Networks (DNNs), also known as deep learning, are the most popular techniques used for Machine Learning today.

Neural Networks

Traditionally, computer programs are built using logical statements which test conditions (if, and, or, etc). But a DNN is different. It is built by training a network of neurons with data alone.

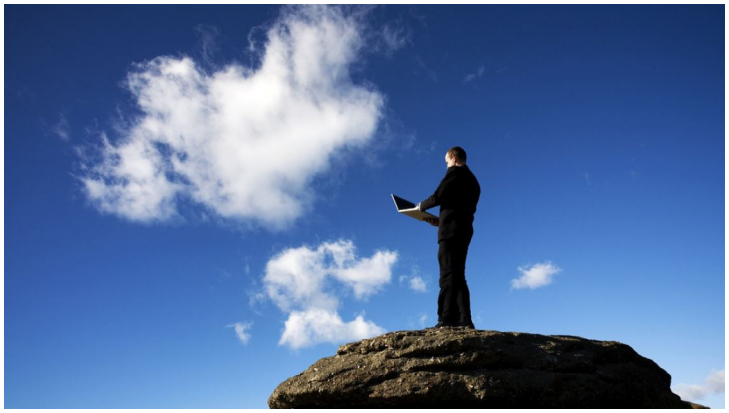
DNN design is complicated, but put simply, there are a set of weights (numbers) between the neurons in the network. Before the training process begins, weights are generally set to random small numbers. During training, the DNN will be shown many examples of inputs and outputs, and each example will help refine the weights to more precise values. The final weights represents what has really been learned by the DNN.

As a result you can then use the network to predict output data given input data with a certain degree of confidence.

Once a network is trained, it is basically a set of nodes, connections, and weights. At this point it is now a static model, one that can be used anywhere needed.

To perform inference on the now static model, you need lots of matrix multiplications and dot product operations. Since these are fundamental mathematical operations, they can be run on a CPU, GPU, or DSP, although the power efficiency may vary.

Cloud



Today, the majority of DNN training and inference happens in the cloud. For example, when you use voice recognition on your smartphone, your voice is recorded by the device and sent up to the cloud for processing on a Machine Learning server. Once the inference processing has occurred, a result is sent back to the smartphone.

The advantage of using the cloud is that the service provider can more easily update the neural network with better models; and deep, complex models can be run on dedicated hardware with less severe power and thermal constraints.

However there are several disadvantages to this approach including time lag, risk of privacy, reliability, and providing enough servers to meet demand.

On-device inference

There are arguments for running inference locally, say on a smartphone, rather than in the cloud. First of all it saves network bandwidth. As these technologies become more ubiquitous there will be a sharp spike in data sent back and forth to the cloud for AI tasks.

Second, it saves power — both on the phone and in the server room — since the phone is no longer using its mobile radios (Wi-Fi or 4G/5G) to send or receive data and a server isn't being used to do the processing.

> Inference done locally delivers quicker results

There is also the issue of latency. If the inference is done locally, then the results will be delivered quicker. Plus there are myriad privacy and security advantages to not having to send personal data up to the cloud.

While the cloud model has allowed ML to enter into the mainstream, the real power of ML will come from the distributed intelligence gained when local devices can work together with cloud servers.

Heterogeneous computing

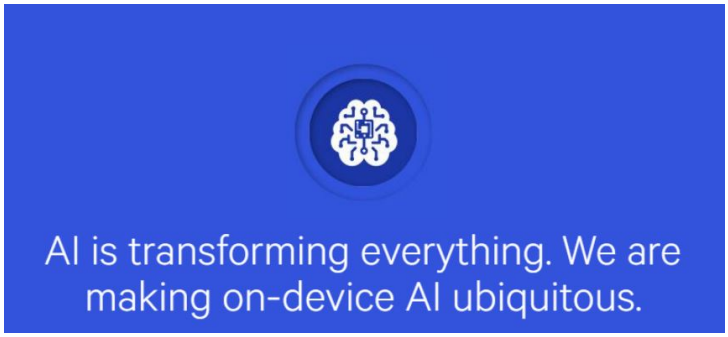
Since DNN inference can be run on different types of processors (CPU, GPU, DSP, etc.), it is ideal for true heterogeneous computing. The fundamental element of heterogeneous computing is the idea that tasks can be performed on different types of hardware, and yield different performance and power efficiency.

For example, Qualcomm offers an Artificial Intelligent Engine (AI Engine) for its premium-tier processors. The hardware, combined with the Qualcomm Neural Processing SDK and other software tools, can run different types of DNNs, in a heterogeneous manner. When presented with a Neural Network built using 8-bit integers (known as INT8 networks), the AI Engine can run that on either the CPU or for better energy efficiency on the DSP. However, if the model uses 16-bit and 32-bit floating point numbers (FP16 & FP32), then the GPU would be a better fit.

> The possibilities for AI augmented smartphone experiences are limitless

The software side of the AI Engine is agnostic in that Qualcomm's tools support all the popular frameworks like Tensorflow and Caffe2, interchange formats like ONNX, as well as Android Oreo's built-in Neural Network API. On top of that there is a specialized library for running DNNs on the Hexagon DSP. This library takes advantage of the Hexagon Vector eXtensions (HVX) that exist in premium-tier Snapdragon processors.

The possibilities for smartphone and smart-home experiences augmented by AI are almost limitless. Improved visual intelligence, improved audio intelligence, and maybe most importantly, improved privacy since all this visual and audio data remains local.



But AI assistance isn't just for smartphone and IoT devices. Some of the most interesting advances are in the auto industry. AI is revolutionizing the future of the car. The long- term goal is to offer high levels of autonomy, however that isn't the only goal. Driver assistance and driver awareness monitoring are some of the fundamental steps towards full autonomy that will drastically increase safety on our roads. Plus, with the advent of better natural user interfaces the overall driving experience will be redefined.

Wrap-up

Regardless of how it is marketed, Artificial Intelligence is redefining our mobile computing experiences, our homes, our cities, our cars, the healthcare industry — just about everything you can think of. The ability for devices to perceive (visually and audibly), infer context, and anticipate our needs allows product creators to offer new and advanced capabilities.

> Machine Learning is redefining our mobile computing experiences

With more of these capabilities running locally, rather than in the cloud, the next generation of AI augmented products will offer better response times and more reliability, while protecting our privacy.

This content brought to you in collaboration with our friends at Qualcomm.